HMM-based speech synthesiser using the LF-model of the glottal source

J.P. Cabral, S. Renals, J. Yamagishi, and K. Richmond. HMM-based speech synthesiser using the LF-model of the glottal source. In Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, pages 4704-4707, May 2011.

HMM-BASED SPEECH SYNTHESISER USING THE LF-MODEL OF THE GLOTTAL SOURCE

João P. Cabral[1,2] *, Steve Renals*[2]*, Junichi Yamagishi*[2] *and Korin Richmond*[2]

[1]School of Computer Science and Informatics, University College Dublin, Ireland
[2]The Centre for Speech Technology Research, University of Edinburgh,UK

joao.cabral@ucd.ie, s.renals@ed.ac.uk, jyamagis@inf.ed.ac.uk, korin@cstr.ed.ac.uk

ABSTRACT

A major factor which causes a deterioration in speech quality in HMM-based speech synthesis is the use of a simple delta pulse signal to generate the excitation of voiced speech. This paper sets out a new approach to using an acoustic glottal source model in HMM-based synthesisers instead of the traditional pulse signal. The goal is to improve speech quality and to better model and transform voice characteristics. We have found the new method decreases buzziness and also improves prosodic modelling. A perceptual evaluation has supported this finding by showing a 55.6% preference for the new system, as against the baseline. This improvement, while not being as significant as we had initially expected, does encourage us to work on developing the proposed speech synthesiser further.

Index Terms— HMM-based Speech Synthesis, LF-Model, Glottal Source Modelling

1. INTRODUCTION

HMM-based speech synthesisers typically generate speech by shaping a spectrally flat excitation with the spectral envelope of speech, e.g. [1]. A simple excitation model consists of using white noise for unvoiced speech and an impulse train for voiced speech. However, this model makes the synthetic speech sound *buzzy* and just allows to control the pitch (through the F_0 parameter). A popular method to reduce the buzziness is to mix the impulse train with a noise component using a multi-band mixed excitation model, e.g. [2]. Recently, other excitation models have been used in statistical speech synthesis which try to better approximate the voiced excitation to the residual calculated using the inverse filtering technique, e.g. [3, 4]. These models can represent more details of the source than the noise. However, they do not model relevant characteristics of the glottal source.

Speech can also be generated by passing a *glottal source model* through a filter representing the vocal tract system. However, the methods to estimate the glottal source and the vocal tract are typically less robust than those to estimate the spectral envelope. Nevertheless, this type of speech model has been successfully used in HMM-based synthesis. For example, the synthesiser in [5] models the glottal source and the vocal tract filter using LPC parameters. During synthesis, the excitation is obtained by transforming a real glottal pulse using the glottal parameters generated by the synthesiser. However, this approach does not allow control over glottal parameters related to voice quality and does not model the correlation between F_0 and the glottal parameters.

In previous work [6], we used an acoustic glottal source model, the Liljencrants-Fant (LF) model [7], in the synthesis part of an HMM-based speech synthesiser. In this system, a selected LF-model signal was passed through a post-filter to obtain a spectrally flat excitation and then speech was generated by shaping the excitation with the spectral envelope.

In this work, we propose another HMM-based speech synthesiser, which generates speech by passing the LF-model signal through the vocal tract filter. The LF-model parameters are trained in the system, which allows the natural variations of the glottal parameters with F_0 to be modelled. In addition, the LF-model parameters can be used to control relevant properties of the glottal pulse shape that are correlated with voice quality, such as breathiness. The vocal tract filter is estimated using the Glottal Spectral Separation (GSS) method [8]. In this paper, we also propose an extension to the GSS synthesis method which consists of mixing the LF-model with a noise component in order to improve speech naturalness further.

2. LILJENCRANTS-FANT MODEL

The Liljencrants-Fant (LF) model [7] is an acoustic model of the glottal source derivative. It can be represented by the following equation:

$$e_{LF}(t) = \begin{cases} E_0 e^{\alpha t}\sin(w_g t), & t_o \le t \le t_e \\ -\frac{E_e}{\epsilon T_a}[e^{-\epsilon(t-t_e)} - e^{-\epsilon(t_c-t_e)}], & t_e < t \le t_c \\ 0, & t_c < t \le T_0 \end{cases} \quad (1)$$

First author is currently supported by the Science Foundation Ireland (Grant 07/CE/I1142). This paper is based on his PhD work supported by Marie Curie Early Stage Training Site EdSST (MEST-CT-2005-020568)

where $w_g = \pi/t_p$. The LF-model is defined by six shape parameters: t_c, t_p, t_e, T_a, T_0, and E_e. The remaining parameters (E_0, ϵ and α) can be calculated using the energy and continuity constraints, which are given by $\int_0^{T_0} e_{LF}(t)dt = 0$ and $e_{LF}(t_e) = e_{LF}(t_e^+) = -E_e$, respectively.

The LF-model is often represented by the first two branches of (1) for simplification. In this case, the instant of complete closure, t_c, is set to the period T_0 in the second branch. In this work, this simplified LF-model was used.

3. BASELINE HMM-BASED SPEECH SYNTHESISER

The baseline statistical speech synthesiser used in this work employed the MATLAB version of the STRAIGHT vocoder (STRAIGHTV40). This system is an implementation of the Nitech-HTS 2005 speech synthesiser [1].

3.1. Analysis

The STRAIGHT analysis method was used to calculate the FFT parameters of the spectral envelope of the short-time speech signal (40 ms long) and aperiodicity parameters (FFT coefficients) measured in the speech spectrum. These parameters were transformed to more suitable features for statistical modelling. The spectral envelope was converted to mel-cepstral coefficients, whereas the aperiodicity measurements were averaged over five frequency bands: 0-1, 1-2, 2-4, 4-6, and 6-8 kHz. Meanwhile, F_0 was estimated using the RAPT algorithm [9].

3.2. Acoustic Modelling

The statistical model was a five-state left-to-right hidden-semi Markov model (HSMM). Both state output density function and state duration were modelled using a single Gaussian distribution. Each observation feature vector consisted of five streams: mel-cepstrum, aperiodicity, $\log F_0$, Δ of $\log F_0$ and Δ^2 of $\log F_0$. The spectrum and aperiodicity parameters were modelled by continuous HMMs, while the last three streams were modelled by multi-space probability distribution HMMs (MSD-HMMs) because F_0 is not defined in unvoiced regions. The spectrum and aperiodicity streams included the static and dynamic features.

The context-dependent models were also clustered using different decision trees for the spectrum, F_0 and duration parameters, since the influence of the contextual factors varies for each of these.

3.3. Synthesis

The STRAIGHT vocoder was used to synthesise speech by convolving a spectrally flat excitation with the spectral envelope of speech (obtained from the mel-cepstrum). For voiced speech, the aperiodicity parameters were used to derive $W_p(w)$ and $W_a(w)$, which are the weighting functions for the spectra of the impulse train (phase manipulated) and white noise respectively. The resulting weighted signals were then added together to obtain the mixed excitation.

4. HMM-BASED SPEECH SYNTHESISER USING LF-MODEL: HTS-LF

The baseline HMM-based speech synthesiser was modified in order to incorporate the LF-model. This system using glottal source modelling is called HTS-LF. The main differences between the two systems are the multi-stream structure of the speech parameter vector and the analysis-synthesis methods.

4.1. Analysis

In the HTS-LF system, the aperiodicity parameters were computed using the STRAIGHT method, whereas the LF-model parameters and the vocal tract spectrum were estimated as in the Glottal Spectral Separation method [8].

4.1.1. LF-model Parameters

The LF-model parameters were estimated from the Linear Prediction (LP) residual, as described in [8]. The residual was computed using the inverse filtering technique with pre-emphasis ($\alpha = 0.97$). Then, the LF-model parameters were calculated for each pitch cycle of the residual, which was delimited by contiguous glottal epochs. The estimation method consisted of fitting the LF-model waveform to the residual using a non-linear optimisation algorithm. The initial estimates of the iterative method were obtained by performing amplitude-based measurements on the residual.

The trajectories of the LF-parameters calculated for an utterance are shown in Figure 1 (a). A strong correlation between the glottal parameters and T_0 can be observed (direct proportion), with the exception of the parameter T_a. Short segments can also be found which show a different pattern of variation with T_0 that is not linear. These may be explained by prosody effects such as accented words and syllable stress

4.1.2. Vocal Tract Spectrum

The speech signal was segmented at 5 ms frame rate into 40 ms long frames. In voiced speech regions, the set of LF-model parameters values associated with each frame $s^j(t)$ was obtained by finding the closest epoch i to the center of $s^j(t)$. These parameters were used to generate one period of the LF-model signal, $e^i_{LF}(t)$. Next, the speech spectrum $S^j(w)$ was divided by the amplitude spectrum of the LF-model signal, $\left|E^i_{LF}(w)\right|$, in order to remove the glottal source model effects. That is, $V^j(w) = S^j(w)/\left|E^i_{LF}(w)\right|$. Finally, the STRAIGHT vocoder was used to calculate the spectral envelope of the signal $V^j(w)$. For unvoiced speech, the spectral parameters were estimated by computing the spectral envelope of $S^j(w)$ using STRAIGHT.

The vocal tract spectrum obtained using the GSS method is expected to be sufficiently smooth, assuming that the LF-model parameter trajectories are smooth enough and that STRAIGHT computes a smooth spectrum. This is considered to be an important characteristic to obtain accurate modelling of the spectrum in the HMM-based speech synthesiser.

4.2. Acoustic Modelling

The statistical modelling part of the HTS-LF system is similar to the baseline system. However, the F_0 parameters vector of the baseline was replaced by the LF-model parameter vector in HTS-LF. The dimension of the LF-model, Δ and Δ^2 streams was set to 5. These streams were modelled by MSD-HMMs using a Gaussian distribution with diagonal covariance matrix for the voiced space. The clustering decision trees for the LF-model parameter streams were built using the same question set and minimum description length criterion as used for clustering the F_0 streams in the baseline system. We assumed the contextual factors most relevant to the LF-parameters were similar to those for the F_0 factors because these parameters are strongly correlated.

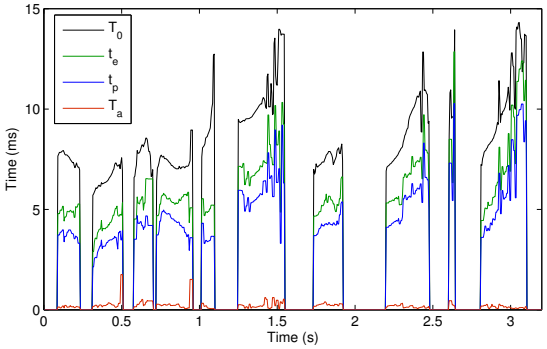
Figure 1 (b) shows that the parameter generation algorithm produces smoother trajectories than those obtained during speech analysis, mainly due to modelling by the HMMs. One advantage of this smoothing effect is attenuation of parameter discontinuities due to estimation errors in analysis.

4.3. Synthesis

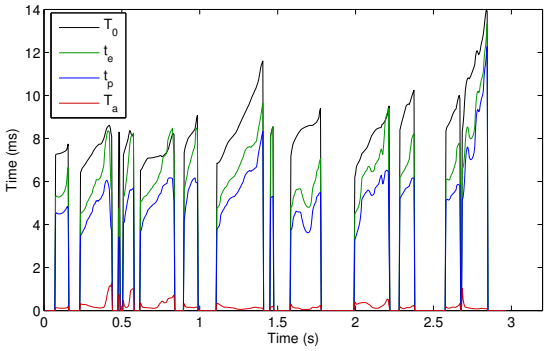
In the GSS method proposed in previous work [8], voiced speech was generated by passing two cycles of the LF-model signal through the vocal tract filter. In this work, the multi-band mixed excitation of STRAIGHT was adapted in order to mix the noise component of the excitation with the LF-model signal. The advantage is to better model the noise component of the speech signal and improve speech naturalness.

Figure 2 shows the flowchart of the synthesis method used by the HTS-LF system. The LF-model signal has a decaying spectrum as it models the spectral tilt of the glottal source. In contrast, the noise spectrum is approximately flat. For this reason, these two signals cannot be mixed using the aperiodicity parameters as in STRAIGHT. In order to overcome this problem, the white noise signal is shaped with the spectral envelope of the LF-model signal before the weighting operation. This shaping is performed in the frequency domain by multiplying the amplitude spectrum of one period of the LF-model signal, $|E_p(w)|$, by the amplitude spectrum of the noise signal, $N(w)$. The resulting noise signal has the same duration as the periodic LF-model signal, $E(w)$, and it is scaled in amplitude by the factor K_n for the two signals to have the same power. The spectrum of the excitation can be represented by:

$$X(w) = E(w)W_p(w) + K_n N(w)|E_p(w)|W_a(w) \qquad (2)$$



(a) Parameters estimated in the analysis.



(b) Parameters generated by the HTS-LF system.

Fig. 1. *Example of trajectories of the LF-model parameters.*

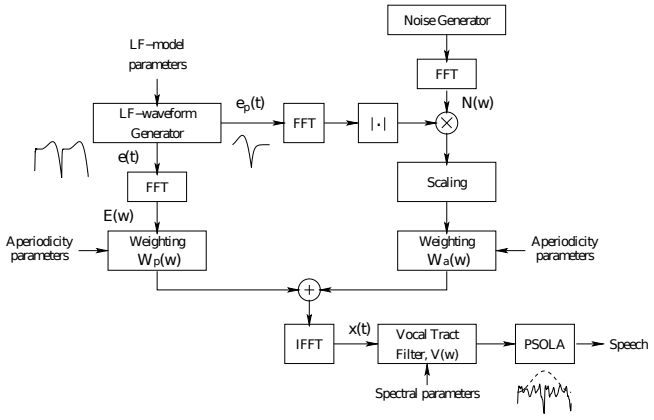


Fig. 2. Block diagram of the speech waveform generation technique used by the HTS-LF system.

Speech is generated by passing the mixed excitation through the vocal tract filter. Finally, the speech frames are concatenated using overlap-and-add with asymmetric windows centered at the instants of maximum excitation. $G(w)$ contains phase information from the LF-model signal which is expected to reduce the buzziness effect of the impulse train.

5. PERCEPTUAL EVALUATION

A forced-choice A-B test was conducted to evaluate the speech quality of the HTS-LF system when compared to the standard HMM-based synthesiser.

5.1. Stimuli

The US English BDL voice (male) was built from the CMU ARCTIC speech database [10] for the two systems. The size of the BDL speech corpus is approximately one hour.

The stimuli consisted of 36 pairs of utterances: 18 utterances synthesised with the two systems, randomly chosen and repeated twice with the order of the samples switched.

5.2. Experiment

The evaluation was conducted via the web. Subjects were asked to listen to the pairs of stimuli and for each pair they had to select the version (A or B) that sounded best. They were able to listen to the files in any order, and as many times as they liked. We also instructed them to make a random choice if they could not decide on the version they preferred.

Students and staff from the University of Edinburgh were asked to perform the evaluation. Fourteen listeners participated in the test, of which six were native speakers of English.

6. RESULTS

The results of the perceptual experiment are shown in Table 1. They are statistically significant with $p <= 0.01$. On average, the HTS-LF system obtained a higher rate of preference. Nevertheless, The results were expected to be even better, as the improvement in the quality of resynthesised natural speech (without modelling) when using the LF-model compared to the impulse train was significantly high in a previous evaluation [8].

From our subjective analysis of the synthetic speech, the "metallic" quality produced by the standard HTS was clearly reduced using the HTS-LF system for some utterances. However, some samples synthesised with the LF-model contained some distortion which might be more perceptually significant than the buzziness characteristic of the standard system. In our opinion, errors in the extraction of the glottal parameters by the HTS-LF system are a possible cause of degradation in speech quality. Also, rapid spectral variations due to the mismatch between the spectral envelope and the vocal tract spectrum at voicing transitions may not have been modelled by the HMMs correctly.

Finally, we also note we have found that prosodic characteristics, e.g. accent position in words, are often better modelled using the HTS-LF system. Examples of the synthesised speech are available at `http://homepages.inf.ed.ac.uk/jscabral/hts-lf-model.html`.

	Baseline	HTS-LF
Mean preference (%)	44.4	55.6
95% Conf. Interv. (%)	[40.1 48.9]	[51.1 59.9]

Table 1. Mean scores and 95% confidence intervals obtained by the two HTS synthesisers in the A-B forced-choice test.

7. CONCLUSIONS

In this work, the LF-model was incorporated into a standard HMM-based speech synthesiser by using the GSS method for analysis-synthesis and adapting the acoustic modelling part to train the glottal parameters.

The proposed HTS-LF system obtained higher preference than an HMM-based speech synthesiser which uses the STRAIGHT vocoder. A great advantage of the HTS-LF system is that it provides control over glottal parameters for voice quality transformations.

There is a good scope for further development of the HTS-LF system. We have been improving the method to estimate the LF-model parameters and studying in detail the causes of speech distortion in this synthesiser.

8. REFERENCES

[1] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, "Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005," *IEICE Trans. Inform. and Systems*, vol. E90-D, pp. 325–333, January 2007.

[2] T. Yoshimura, K. Tokuda, T. Masukom, T.and Kobayashi, and T. Kitamura, "Mixed excitation for HMM-based speech synthesis," in *Proc. of EUROSPEECH*, Aalborg, September 2001.

[3] R. Maia, T. Toda, H. Zen, Y. Nankaku, and K. Tokuda, "A trainable excitation model for HMM-based speech synthesis," in *Proc. of INTERSPEECH*, Antwerp, August 2007.

[4] T. Drugman, G. Wilfart, and T. Dutoit, "A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis," in *Proc. of INTERSPEECH*, Brighton, September 2009.

[5] T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku, "HMM-based Finnish text-to-speech system utilizing glottal inverse filtering," in *Proc. of INTERSPEECH*, Brisbane, 2008.

[6] J. Cabral, S. Renals, K. Richmond, and J. Yamagishi, "An HMM-based speech synthesiser using Glottal-Post Filtering," in *Proc. of the 7th SSW*, Japan, September 2010.

[7] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," STL-QPSR, KTH, Stockholm, 1985.

[8] J. Cabral, S. Renals, K. Richmond, and J. Yamagishi, "Glottal spectral separation for parametric speech synthesis," in *Proc. of the INTERSPEECH*, Brisbane, 2008.

[9] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. 1995, pp. 495–518, Elsevier Science.

[10] J. Kominek and A. Black, "The CMU Arctic speech databases," in *Proc. of 5th SSW*, Pittsburgh, June 2004.